FernUniversität
Gesamthochschule in Hagen

Universitätsrechenzentrum

Thomas Feuerstack
Abt. Wiss. Anwendungen

Über die Kunst und das Vergnügen, Texte und Graphiken zu scannen

(Kurzeinführung zu *HP DeskScan* und *Omnipage Pro*)

B/062/0007

© FernUniversität, Juli 2000

Inhaltsverzeichnis

I Grundlagen **5**

1 Vorbemerkung **7**

1.1 Vorkenntnisse 7

1.2 Begriffserklärung 7

1.3 Hardware 8

1.4 Software 9

II Scannen von Graphiken **11**

2 Einführung in *DeskScan II* **13**

2.1 Was ist Scannen? 13

2.2 Und wie geht das? 13

III Scannen von Texten **21**

3 Einführung in *Omnipage Pro* **23**

3.1 Die Kunst der OCR 23

3.2 Die Verwendung von *Omnipage Pro* 24

3.3 Weitere Tipps und Kniffe 30

3.4 Benutzung des Scanners HP 9100C Digital Sender . . . 30

I Grundlagen

1 Vorbemerkung

Vor den Erfolg haben die Götter den Schweiß gesetzt! Dies gilt insbesondere für die Arbeit im Graphik- oder Layoutbereich, den Sie an dieser Stelle über den Seiteneingang des *Scannens* betreten. Sicherlich, diese Aussage vorweg mag für Sie ziemlich frustrierend klingen, wo doch die Werbung ständig verspricht moderne Software erledigt alles für Sie von allein; jedoch gute Ergebnisse verlangen immer noch ein hohes Quantum an Mitarbeit; Software kann hier nur ein Hilfsmittel sein – wie Sie diese Hilfsmittel jedoch möglichst sinnvoll nutzen, dafür mag die vorliegende Broschüre eine erste Anregung sein.

Viel Vergnügen bei der Lektüre.

1.1 Vorkenntnisse

Sicherlich werden Sie jetzt erstaunt aufblicken und nach dem Sinn einer Einführungsbroschüre fragen, in welcher von Vorkenntnissen die Rede ist. Nun, für die Bedienung der beschriebenen Produkte *DeskScan II* und *Omnipage Pro* werden an dieser Stelle keine Erfahrungen Ihrerseits vorausgesetzt, auch wenn Vorkenntnisse im Bereich der allgemeinen grafischen Datenverarbeitung durchaus hilfreich sind.[1]

Ein anderer Punkt ist die Bedienung des grundlegenden Betriebssystems beider Produkte. Ich gehe im Laufe dieser Broschüre davon aus, dass Sie Grundkenntnisse im Betriebssystem *Windows*[2] haben; dass beispielsweise ein Dialogfeld immer bestätigt oder abgebrochen werden muß, habe ich nicht immer explizit erwähnt. Ebenso setze ich voraus, daß Sie wissen, wie Fenster geöffnet, verschoben und geschlossen werden, den Datei-Manager bedienen können. . . das Einmaleins von Windows eben.

Falls dies für Sie nicht zutrifft, eignen Sie sich bitte die grundlegenden *Windows*-Arbeitsweisen an, bevor Sie mit der eigentlichen Scanner-Arbeit beginnen. Die Möglichkeit dazu haben Sie u.a. durch den Besuch eines vom Universitätsrechenzentrum veranstalteten Einführungsseminars.

1.2 Begriffserklärung

Im weiteren Verlauf dieser Einführung werden häufig sehr themenbezogene Begriffe auftauchen, die dem Laien wahrscheinlich nicht geläufig sein werden. Aus diesem Grunde möchte ich mich als erstes mit Ihnen

[1] Ansonsten habe ich versucht, die wichtigsten fachspezifischen Begriffe im nächsten Kapitel kurz zu erläutern.

[2] Mit *Windows* sind prinzipiell die Versionen *95, 98, NT* und *2000* gemeint.

auf ein gemeinsames Vokabular einigen, damit wir hinterher möglichst wenig Missverständnisse ausräumen müssen.

Scanner sind Geräte welche die Aufgabe haben, vorliegende Bilder[3] zu digitalisieren, so dass diese mit einer zusätzlichen Software weiterverarbeitet werden können. In seiner prinzipiellen Arbeitsweise besitzt ein Scanner durchaus Ähnlichkeiten mit dem wesentlich häufiger anzutreffendem Fotokopierer. Und um in einem Atemzug mit einem verbreiteten Vorurteil aufzuräumen: es gibt weder spezielle Graphik- noch Text-Scanner; in welchem Format Ihr Bild letztendlich abgespeichert wird hängt ausschließlich von der Software ab, die mit dem Scanner zusammenarbeitet.

Scannen bezeichnet den eigentlichen Vorgang der Digitalisierung. Hierbei wird das vorliegende Bild vom Scanner abgetastet und in einzelne Rasterpunkte zerlegt. Im Anschluß daran kann das Raster mittels eines graphischen Editors, z.B. *PC-Paintbrush* oder dem *Adobe's Photoshop* weiter bearbeitet, bzw. durch eine geeignete OCR-Software in einen maschinenlesbaren Text umgewandelt werden.

OCR ist die Abkürzung für *Optical Character Recognition.* In diesem speziellen Fall wird versucht, aus einem durch einen Scanner erzeugtem Raster wieder einen maschinenlesbaren Text herzustellen, der dann durch herkömmliche Textverarbeitungssoftware wie z.B. *Word für Windows* oder LaTeX weiterverarbeitet werden kann.

1.3 Hardware

Am Universitätsrechenzentrum der FernUniversität wird das Scannen jeglicher Art z.Zt. mit folgender Hardware realisiert.

- Ein Scanner *Hewlett & Packard Scanjet 4c* mit maximaler Auflösung von 600dpi (2400dpi enhanced) und Darstellungsmöglichkeit von 16,7 Millionen Farben bzw. 256 Graustufen.
- Ein zugehöriger Einzelblatteinzug für ca. 50 Blatt bis hin zur Größe DIN A4.

Tipp!

- *Neu* ist an dieser Stelle die Verwendung des in der URZ-Beratung (Raum A002) befindlichen Netzwerk-Scanners *HP 9100C Digital Sender*, zum Einlesen umfangreicherer Dokumente. Detaillierte Hinweise zur Benutzung dieses Scanners finden Sie am Ende dieser Broschüre.

Bedienung der Hardware

Die Bedienung der Hardware sollte eigentlich kein Problem darstellen, es seien an dieser Stelle lediglich ein paar Eigenheiten des Scanners auf-

[3] *Bilder* sind in diesem Zusammenhang Darstellungen, die auf einem Sekundärträger vorliegen. Während der Sekundärträger zu 99% aus Papier besteht (das restliche Prozent kann durch Pappe, Plastik oder bedruckte T-Shirts (s. Titelbild) abgedeckt werden), kann es sich bei dem darauf befindlichen *Bild* sowohl um eine „echte" Graphik, als auch um einen Text handeln.

gezählt, um einen reibungslosen Ablauf zu gewährleisten.

1. Schalten Sie den Scanner unbedingt ein *bevor* Sie den angeschlossenen Rechner aktivieren oder booten. Der Schalter des Scanner befindet sich einigermaßen versteckt auf der rechten Seite im hinteren Bereich.

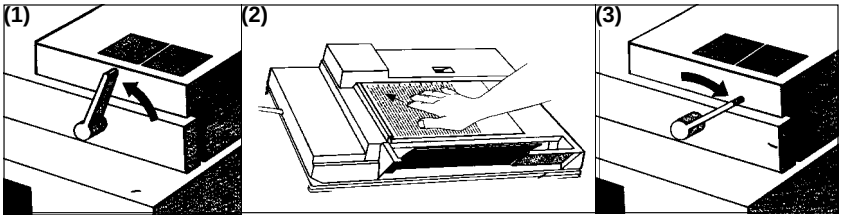

Abbildung 1.1: Benutzung des Einzelblatteinzugs in drei Stufen.

2. Falls Sie den Einzelblatteinzug benutzen, gehen Sie bitte so vor, wie es die Abbildung 1.1 zeigt, d.h. schieben Sie den im vorderen Teil befindlichen *grünen Hebel* nach links (1) legen Sie Ihre Seiten ein (2) (dabei liegt die erste Seite *oben*, so dass Sie sie lesen können) und schieben Sie diese bis zum Widerstand. Legen Sie den grünen Hebel jetzt wieder nach rechts um (3).

3. In Fällen in denen der Einzelblatteinzug ungeeignet erscheint, z.B. beim Scannen aus Büchern oder Illustrierten, können Sie ihn wie einen Deckel hochklappen. Legen Sie die zu scannende Seite mit dem Bild nach unten auf die Glasscheibe, so dass der Seitenkopf von Ihnen wegzeigt. Richten Sie die Seite nun an der oberen rechten Ecke aus.

! → *Beim Scannen aus Büchern reicht es vollkommen aus, wenn der nun als Deckel fungierende Einzelblatteinzug locker auf dem Buch aufliegt. Versuchen Sie auf keinen Fall den Deckel zu beschweren oder gar mit der Hand zu drücken. (Bruchgefahr!)*

   Entfernen Sie nach dem Scannen das Buch von der Glasscheibe des Scanners, und senken Sie den Einzelblatteinzug herab, bis er hörbar einrastet.

1.4 Software

Nachdem im letzten Kapitel die (hoffentlich!) pflegliche Behandlung der Hardware angesprochen wurde, müssen Sie als letzte Hürde lediglich wissen welche Software Sie zur Nutzung des Scanners verwenden können, bzw. wo sich diese auf dem Rechner versteckt. Momentan sind folgende, zu unserem Scanner kompatible Softwareprodukte installiert.

- *HP DeskScan II* zum Einscannen von Graphiken und deren Speicherung in einem softwareübergreifendem Rasterformat.
- *Omnipage Pro* Version 9.0 zum Einscannen von Texten und deren Konvertierung in ein textverarbeitungsübliches Format.

II Scannen von Graphiken

2 Einführung in *DeskScan II*

2.1 Was ist Scannen?

> Das Scannen eines Bildes entspricht dem Ablichten eines Bildes mit einer Kamera. [...] Im Gegensatz zu einer Kamera können Sie mit der Scanner-Software Bildinformationen ändern, bevor Sie es im Computer speichern oder in ein Desktop-Publishing- oder Bildbearbeitungsprogramm exportieren. Sie können beispielsweise das Bild vergrößern oder verkleinern, den Kontrast ändern, die Helligkeit einstellen und das Bild auf andere Weise umgestalten, so dass das endgültige Bild u.U. sogar besser aussieht als das Original.
>
> (aus *HP DeskScan II Benutzerhandbuch*)

2.2 Und wie geht das?

Nun, hoffentlich ziemlich einfach. Wir vom Universitätsrechenzentrum haben die *DeskScan II*-Software so installiert, dass in einem Großteil der Fälle nachträgliche Retuschierungen Ihres eingescannten Bildes auf ein Minimum begrenzt werden können.

Nichtsdestotrotz besteht natürlich alternativ die Möglichkeit jede der von uns getroffenen Voreinstellungen mit eigenen Werten zu überschreiben. Möglichkeiten für die Lösung spezieller Probleme werden im weiteren Verlauf der Broschüre noch vorgestellt.

2.2.1 Der erste Versuch

Nachdem Sie Scanner und Rechner eingeschaltet und *Windows* gestartet haben, starten Sie *DeskScan* über die Menüreihenfolge
`Start - Programme - HP DeskScan II - HP DeskScan II`.

Abbildung 2.1 auf der nächsten Seite

Im Anschluss daran sollten Sie das *DeskScan II*-Hauptfenster auf dem Bildschirm erkennen.

- Legen Sie jetzt die zu scannende Graphik in den Scanner bzw. in den Einzelblatteinzug ein (s.a. Kapitel 1.3)

Abbildung 2.2 auf der nächsten Seite

- Klicken Sie nun auf den Knopf `Voransicht`. Die im Scanner befindliche Seite wird abgetastet, bzw. die erste im Einzelblatteinzug befindliche Seite wird eingezogen. *DeskScan II* stellt die gescannte Seite im rechten Teil des Fensters dar und umgibt den sogenannten „Bild-Bereich“ mit einem Rahmen.

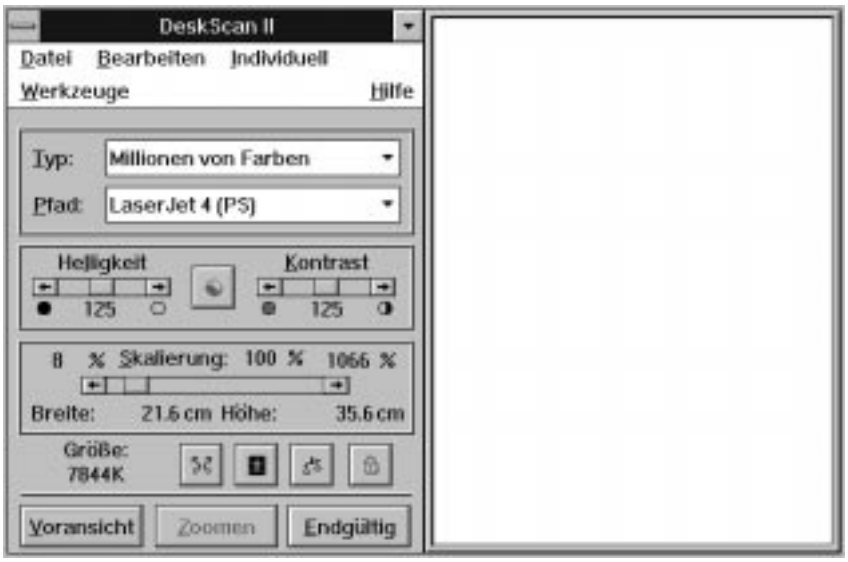

Abbildung 2.1: *DeskScan II* – Hauptfenster.

! → Abbildung 2.2

Neben dem Versuch die gewünschte Graphik auf dem Papier herauszuheben, sollte *DeskScan II* auch den Typ des Bildes erkannt und automatisch eingestellt haben. Ist dies nicht der Fall, sollten Sie die dafür nötigen Optionen über das Menü `Bearbeiten - Optionen` einschalten.

- Positionieren Sie jetzt mit Hilfe der Maus den Rahmen über dem zu scannenden Objekt der aktuellen Seite.

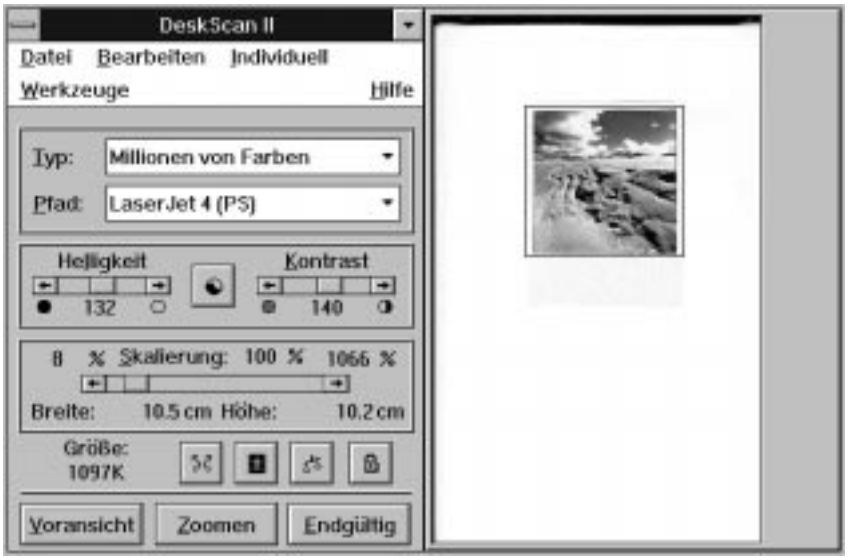

Abbildung 2.2: Das Hauptfenster nach `Voransicht`.

- Klicken Sie auf die Schaltfläche `Zoomen`. Der ausgewählte Bereich wird erneut gescannt und wiederum auf dem Bildschirm angezeigt. Korrigieren Sie, wieder durch Einsatz der Maus, falls nötig den eingestellten Rahmen.

- Klicken Sie auf die Schaltfläche `Endgültig`. Sie erhalten ein Fenster (Abbildung 2.3), in welchem Sie den Namen und den verwendeten Dateityp[1] eintragen. Klicken Sie hier auf `OK` um die Verarbeitung zu beenden.

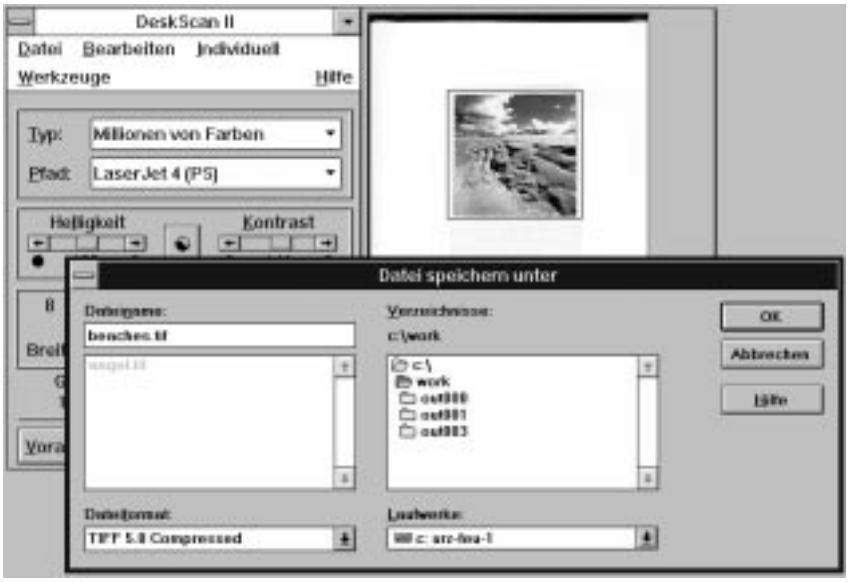

Abbildung 2.3: Das Sicherungsfenster nach Betätigung der Schaltfläche `Endgültig`.

An dieser Stelle haben Sie – wenn auch stark automatisiert – eine komplette Graphik gescannt und abgespeichert. Abweichungen von dieser Verfahrensweise müssen Sie nur vornehmen, falls Sie gezielt auf Charakteristika wie z.B. Kontrast oder Farbtiefe Einfluß nehmen wollen. Einige Werkzeuge dazu werden Sie in den folgenden Kapiteln kennenlernen.

2.2.2 Voreinstellungen

Da Sie sich den Scanner-PC mit mehreren MitbenutzerInnen teilen müssen, ist es primär sinnvoll erst einmal die Standardbelegungen vorzustellen, bevor diese verändert werden. Wählen Sie dazu aus der Menüleiste den Punkt `Bearbeiten` und verzweigen Sie zum Menüpunkt `Optionen...` Sie sollten danach das folgende Fenster (2.4) mit den dargestellten Voreinstellungen sehen. Zum besseren Verständnis werden an dieser Stelle die wichtigsten Menüpunkte erläutert.

Farbvoransicht : stellt beim Vorscannen das gerasterte Bild farbig dar – natürlich nur, wenn das Bild auch in Farbe vorliegt.

Aktual. Voransicht : aktualisiert bei jedem weiteren Vorscannen jeweils die Voransicht. Beide Optionen beeinflussen den genutzten Hauptspeicher des Sys-

[1] *DeskScan II* bietet Ihnen als Dateityp mehrere Metafile-Formate an. Das Metafile-Format benötigen Sie, um die Graphik später mit dem von Ihnen gewünschten Produkt weiterverarbeiten zu können. Sofern Sie an dieser Stelle unsicher sein sollten, wählen Sie eines von den TIFF- oder PCX-Formaten, die mit den gängigsten Softwarelösungen weiter verarbeitet werden können.

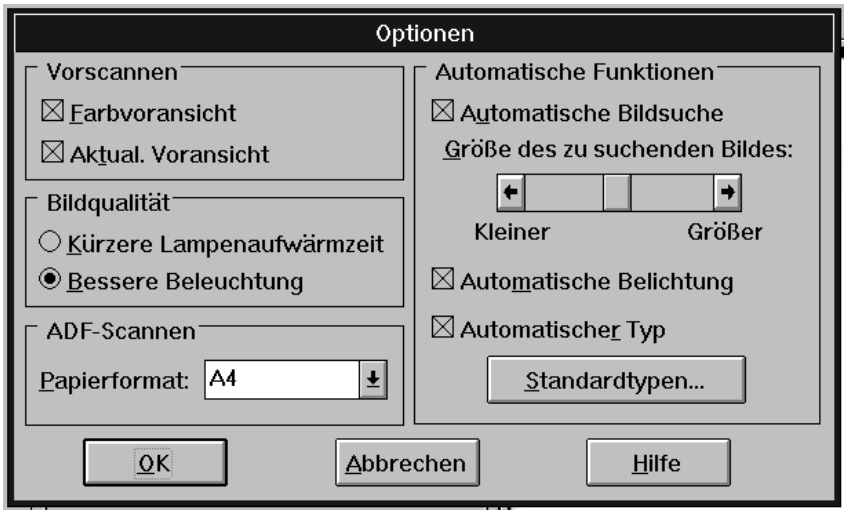

Abbildung 2.4: Die vom Universitätsrechenzentrum gesetzten Voreinstellungen.

tems, der jedoch auf dem Scanner-PC `Karin` in ausreichendem Maße vorhanden sein sollte.

Kürzere Lampenaufwärmzeit : erzielt verkürzte Scan-Zeiten liefert dafür jedoch nicht immer eine optimale Belichtung.

Bessere Beleuchtung : erzeugt ein optimal ausgeleuchtetes Bild. Der Nachteil ist jedoch, dass die Scanner-Lampe vor dem Scannen 8-10 Sekunden aufgewärmt wird. Die beiden letzten Optionen sind alternativ verwendbar.

Papierformat : definiert die Größe der im Einzelblatteinzug[2] befindlichen Vorlage(n). Diese sollte standardmäßig auf A4 gesetzt sein.

Automatische Bildsuche : versucht auf der Vorlage einen „Bildbereich“ zu erkennen und einzurahmen.

Automatische Belichtung : setzt Helligkeits- und Kontrastattribute für den Scan-Bereich.

Automatischer Typ : ordnet Ihrem Bild einen Scantyp zu, dies kann sich u.a. auf Farbe, Rasterung u.v.a.m. erstrecken.

! → Selbstverständlich dürfen Sie alle diese Voreinstellungen mit eigenen Werten überschreiben. Beachten Sie jedoch, dass ein wahlloses Ausprobieren in der Regel eher eine Verschlechterung des Ergebnisses als eine Verbesserung nach sich zieht.

Nachdem wir diesen kurzen Optionen-Ausflug beendet haben, kehren wir zurück ins Haupt-Fenster, zu einer präzisen Beschreibung der dortigen Funktionen.

[2] ADF = Automatic Document Feeder

2.2.3 Optionen im Hauptfenster

Über die Einstellmöglichkeiten im Hauptfenster haben Sie vielfältige Möglichkeiten den Scan-Vorgang zu beeinflussen.

Typ : In diesem Feld erscheint der von *DeskScan II* erkannte bzw. vermutete Bildtyp Ihrer Vorlage. Haben Sie die Option *Automatischer Typ* aktiviert, werden Sie diesen in den seltensten Fällen ändern müssen.

→ Automatischer Typ

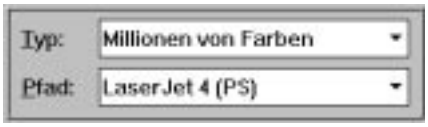

Abbildung 2.5: Die Felder Typ und Pfad.

Pfad : gibt an für welches Ausgabegerät die Vorlage eingescannt wird. Wählen Sie zwischen einem der dort aufgeführten Drucker bzw. *Bildschirm*.

Beide Optionen regulieren letztendlich die *Auflösung* in welcher Ihr Bild gescannt wird. So wird ein Bild mit dem Pfad *Bildschirm* in einer geringeren Auflösung eingescannt als dies z.B. für den Pfad *LaserJet 4 (PS)* der Fall sein würde.

Beachten Sie jedoch: Je *geringer* die verwendete Auflösung ist, desto deutlicher erscheint Ihr Bild gerastert. Je *höher* die verwendete Auflösung ist, desto mehr Plattenspeicherplatz[3] verbraucht Ihr Bild.

Tipp! Bilder, die im World Wide Web plaziert werden sollen, können mit einer Auflösung von 75dpi eingescannt werden, da Bildschirme in aller Regel kaum höhere Auflösungen erreichen; es genügt daher diese mit dem Pfad *Bildschirm* zu versehen. Sollte das Ergebnis Ihren Ansprüchen nicht genügen bringt zumeist ein nachträgliches *Einschärfen* (s.a. Kapitel 2.2.4 auf der nächsten Seite) den gewünschten Erfolg.

Helligkeit/Kontrast : Über beide Schieberegler können Sie die Belichtung des Scanners regeln, eine tiefergehende Erklärung ist an dieser Stelle wohl nicht nötig. Interessant ist jedoch der Button *Automatische Belichtung* der sich dazwischen befindet. Dieser stellt Helligkeit und Kontrast entsprechend des (in Kapitel 2.2.1 gezogenen) Auswahlbereichs automatisch ein.

! → Um ein optimales Ergebnis zu erzielen sollte nach jeder Änderung des Auswahlbereiches die *automatische Belichtung* erneut durchgeführt werden.

[3] Der benutzte Plattenspeicher steht natürlich in direkter Abhängigkeit zu der Zeit, die ein Programm braucht, um ein Bild anzuzeigen. Nehmen Sie als Beispiel ein gescanntes Photo mit einer Auflösung von 600dpi, so wird ein Programm wie *HP Paintbrush* eine gute Minute brauchen um es zu laden. Die gleiche Funktion hätte bei einer Auflösung von 75dpi lediglich 10 Sekunden gedauert.

Abbildung 2.6: Die Schieberegler Helligkeit und Kontrast.

Abbildung 2.7: Skalierung und weitere Möglichkeiten der Bildbeeinflussung.

Skalierung : definiert die ultimative Größe Ihres Bildes. Standardmäßig erfolgt eine 1:1 Skalierung, d.h. Ihr Bild wird in der gleichen Größe gespeichert in welcher es als Original vorliegt. Scannen Sie Ihr Bild grundsätzlich in den Dimensionen ein, in denen Sie es benötigen; eine nachträgliche Skalierung über externe Programme führt meist zu negativen Begleiterscheinungen wie auftretenden Moire-Mustern[4] o.Ä.

Skalierung, ungleichmäßig : Eine ungleichmäßige Skalierung, d.h. eine Verzerrung des Bildes, erhalten Sie, indem Sie auf die Schaltfläche klicken. Anstelle einer Skalierungsskala sehen Sie nun derer zwei. Ein erneuter Druck auf den Button schaltet wieder auf die gewohnte gleichmäßige Skalierung zurück.

Sperren : Durch Druck auf den Button können Sie die momentane Bildgröße sperren und auf weitere zu scannende Bilder übertragen. Dies ist insbesondere dann von Nutzen, wenn Sie mehrere Bilder in der gleichen Größe einlesen wollen.

Spiegelung : Eine vertikale Spiegelung Ihres gescannten Bildes erhalten Sie durch die Betätigung des Buttons .

Invertierung : Eine inverse Darstellung erzielen Sie durch den Button .

2.2.4 Weitere Werkzeuge

Weitere Werkzeuge stehen Ihnen für den Fall zur Verfügung, dass Sie bei sehr schwierig einzuscannenden Graphiken (z.B. bei einer sehr groben

[4] Mein besonderer Dank gilt an dieser Stelle meiner Kollegin Manuela Jürgens, die bei der Korrekturlesung anmerkte, dass sich wahrscheinlich nicht jede LeserIn etwas unter dem Begriff *Moire-Muster* vorstellen kann. Also: Bei einer überdimensionalen Skalierung der gescannten Graphik mittels eines externen Programms, tritt irgendwann der Zustand ein, dass sich einzelne Pixel nicht mehr stauchen bzw. strecken lassen. Dies resultiert i.A. in einer hell/dunklen rechteckigen Zerlegung des Bildes, den sog. *Moire*s.

Rasterung der Vorlage) Korrekturen vornehmen wollen. Sie alle vorzustellen würde sicherlich den Rahmen einer Einführung sprengen, deshalb sollen nur einige wenige exemplarisch erwähnt werden.

Bildschärfe : Im Menü `Individuell - Bildtyp...` finden Sie den Menüpunkt `Bildschärfe`, mit dessen Hilfe Sie feine, individuelle Details hervorheben können. `Bildschärfe` kann für folgende Aufgaben verwendet werden:

- mit niedriger Auflösung gescannte Zeichnungen und Fotos verbessern
- feine Details in Fotos verbessern, damit sie besser hervortreten als im Original
- feine Details in Zeichnungen hervorheben

! → `Bildschärfe` sollte nicht verwendet werden, sofern das Bild „kleine Fehler“ enthält, bzw. bereits gerastert (beispielsweise Bilder aus Zeitungen oder Illustrierten) vorliegt.

Farbanpassung : Da der *ScanJet 4c* auch eine Farbkomponente besitzt, haben Sie selbstverständlich die Möglichkeit, die Farbanteile Ihres Bildes während des Scan-Vorgangs zu schmälern, bzw. zu intensivieren. Wählen Sie hierzu aus dem Menü `Werkzeuge` die Option `Farbanpassung`. Sie erhalten ein externes Fenster mit Schieberegler, über den Sie die Farbsättigung regulieren können.

Alle vorgenommenen Änderungen werden direkt durch die *aktualisierte Voransicht* sichtbar.

Verstärkung : Neben der Farbanpassung besitzen Sie im gleichen Menü `Werkzeuge` noch die Möglichkeit, helle Bildstellen, Zwischentöne und dunkle Bildstellen separat zu akzentuieren.

! → Dieses Werkzeug sollte nur von erfahrenen Scanner-Benutzern eingesetzt werden.

III Scannen von Texten

3 Einführung in *Omnipage Pro*

3.1 Die Kunst der OCR

Wie bereits erwähnt handelt es sich beim Prozess der *Optical Character Recognition* (oder kurz: OCR) darum, eine eingescannte Seite so umzuwandeln, dass das Ergebnis wieder für eine Textverarbeitung, wie beispielsweise Word für Windows zu gebrauchen ist.

Dies ist, technisch betrachtet, heutzutage kein Problem mehr. Sofern Sie irgendwann einmal einen Scanner gekauft haben, werden Sie mit ziemlicher Sicherheit auch eine OCR-Software dazu mitgeliefert bekommen haben. Es handelt sich hierbei jedoch größtenteils um eine funktional eingeschränkte oder abgespeckte Demo-Version.

Doch auch wenn Sie etwas tiefer in die Geldbörse gegriffen haben, um eine OCR-Vollversion (wie z.B. das hier vorgestellte *Omnipage Pro*) zu benutzen, erfüllt das erzeugte Ergebnis häufig nicht die hochgesteckten eigenen Erwartungen. Wo sind die Gründe hierfür zu suchen?

Die meisten Softwarehersteller versprechen Ihnen durch die Verwendung ihrer Produkte eine *Erkennungsquote* von mindestens 99%. Diese Aussage ist zwar theoretisch korrekt, doch Erkennungsquoten dieser Grössenordnung werden in aller Regel nur mit extrem gut vorbereiten Vorlagen erzielt. Achten Sie daher, bevor Sie sich mit der eigentlichen OCR beschäftigen, auf die folgenden Punkte damit Ihre ersten Versuche nicht direkt in einer Enttäuschung enden:

Vorlagen : Die einzuscannende Vorlage sollte möglichst sauber sein. Dies bedeutet unter anderem, dass sie frei von jeglichen Schmutzeffekten, beispielsweise Flecken durch die Verwendung eines Kopierers, sein sollte. Auch sollte der auf der Vorlage befindliche *Seitenspiegel*[1] nach Möglichkeit gerade sein.

Erkannte Zeichen : Häufig verwendete Schriftarten, z.B. Times Roman, Arial oder Helvetica, werden eigentlich problemlos erkannt. Schwierig wird es jedoch bei der Verwendung sogenannter Exoten wie Frakturen oder Handschriften. Gute OCR-Programme besitzen in solchen Fällen die Möglichkeit, diese Schriftarten zu „trainieren“.

Mathematische Formeln : stellen für jede OCR-Software ein Problem dar und werden daher ignoriert.

[1] Unter einem Seitenspiegel versteht der Fachmann die Ausrichtung der auf der Seite befindlichen Textzeilen. Gerade bei einer kopierten Vorlage besteht die große Gefahr des Verrutschens.

Formaterkennung : Auch hier gibt es häufig lange Gesichter. Gewünscht wird natürlich, dass die auf der Vorlage befindlichen Textformate, also Schriftart, -grösse, und -schnitt, ebenso wie die Seitenaufteilung (Spaltensatz!) übernommen werden.

Die meisten OCR-Systeme geben sich in diesem Fall zwar grosse Mühe, doch das Ergebnis ist trotzdem stark nacharbeitungsbedürftig. Denken Sie daran, dass Sie die eingescannten Texte in aller Regel in andere Dokumente integrieren wollen, und verzichten Sie deshalb besser auf eine komplett übernommene Formatierung.

Alles in allem sollen Ihnen die letzten Punkte im wesentlichen eines verdeutlichen: Ein eingescannter Text wird *nie* haargenau wie die verwendete Vorlage aussehen! Machen Sie sich daher stets auf ein gewisses Maß an Nachbearbeitung gefasst. Je geringer die Erwartungen sind, die Sie an Ihre OCR-Software stellen, desto kleiner ist letztendlich die Enttäuschung!

3.2 Die Verwendung von *Omnipage Pro*

Die folgenden Schritte sollen Ihnen zeigen, wie Sie möglichst schnell an brauchbare Dokumente mit Hilfe von *Omnipage Pro* gelangen, ohne sich intensiv in das System einarbeiten zu müssen.

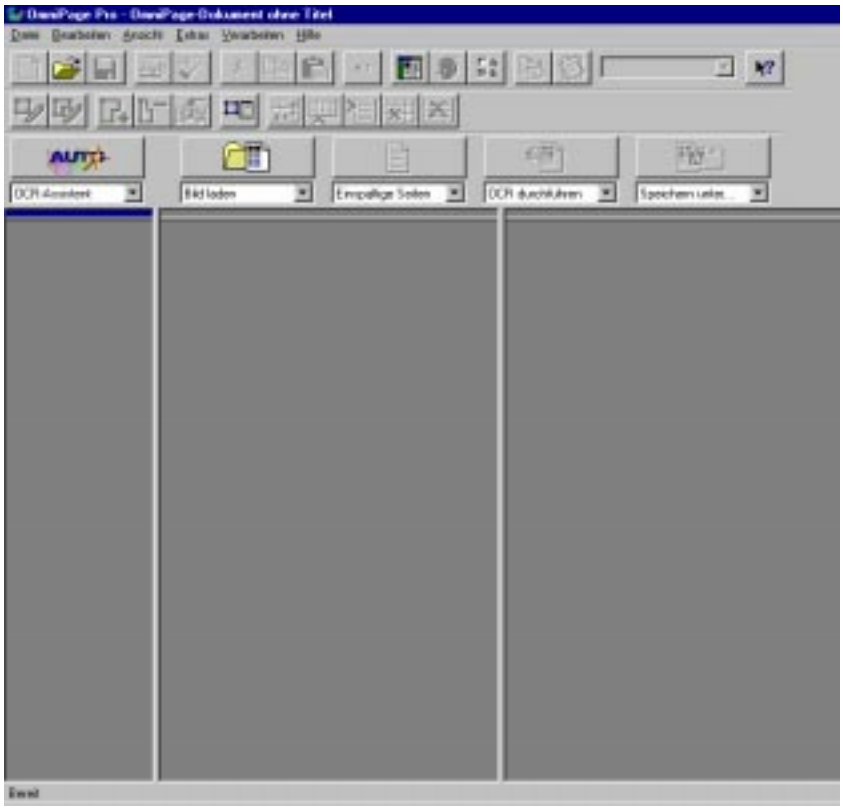

Abbildung 3.1: Das Eingangsfenster von *Omnipage Pro*

Starten Sie *Omnipage Pro* über die Menüreihenfolge `Start - Programme - Omnipage 9.0 Prof. - Omnipage Pro 9.0`[2]

Abbildung 3.1 auf der vorherigen Seite

Sie sollten daraufhin Omnipage's Eingangsfenster auf dem Bildschirm sehen.

Bevor Sie mit der eigentlichen Erkennung beginnen, müssen Sie jetzt vorab den Scanner konfigurieren – schließlich weiß Omnipage zu diesem Zeitpunkt noch nicht, ob Ihre im Scanner befindlichen Vorlagen ein- bzw. doppelseitig bedruckt sind.

Abbildung 3.2

Wählen Sie daher als erstes aus dem Menü `Extras` den Eintrag `Optionen...` und klicken Sie in dem erscheinenden Fenster auf die Karteikarte `Scanner`. Achten Sie darauf, dass die Option Doppelseitig bedruckt gemäß Ihrer Vorlage inaktiv bzw. aktiviert ist.

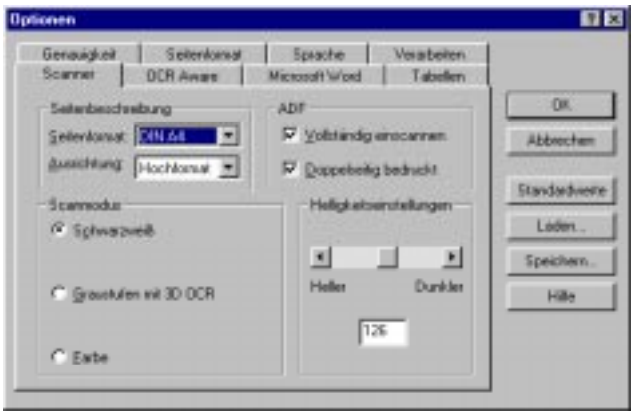

Abbildung 3.2: Konfiguration des Einzelblatteinzugs.

! → Beim Einscannen doppelseitig bedruckter Vorlagen zieht der Scanner zuerst alle Frontseiten ein. Nach Aufforderung durch Omnipage drehen Sie den Stapel um, d.h. die letzte Seite befindet sich jetzt oben, und legen ihn erneut in den Einzelblatteinzug.[3] Bestätigen Sie *erst danach* die Aufforderung mit dem Scannen fortzufahren.

Nach der Konfiguration des Scanners können Sie nun mit der eigentlichen Verarbeitung beginnen. Wählen Sie, sofern er nicht bereits voreingestellt ist, aus der Auswahlliste über der linken Spalte den Eintrag `OCR-Assistent`, so dass Sie den Button AUTO sehen.

Die weitere Arbeit überlassen wir nun günstigerweise Omnipage's OCR-Assistenten, der uns in sechs Schritten zum fertigen Dokument führen wird.

Starten Sie den Assistenten durch Klick auf AUTO.

[2] Das Verzeichnis `Omnipage 9.0 Prof.` scheint eine Eigenart des Rechners `Karin` zu sein. Sollten Sie die Software auf Ihrem eigenen Rechner benutzen, versteckt sie sich wahrscheinlich im Ordner `Caere-Anwendungen`.

[3] Omnipage sortiert die eingelesenen Seiten automatisch in der richtigen Reihenfolge!

Schritt 1 – Auswahl der Vorlage

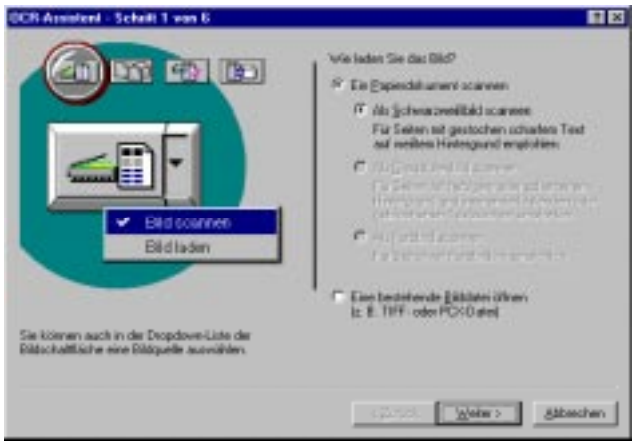

Abbildung 3.3: Schritt 1

Abbildung 3.3 In diesem Fenster legen Sie fest, wo sich Ihre Dokumenten-Vorlage befindet. Sofern sie bereits im Einzelblatteinzug des Scanners vorliegt, wählen Sie die Option Ein Papierdokument scannen.

Sollten Sie Ihre Vorlage bereits anderweitig eingescannt und auf der Festplatte gespeichert haben, laden Sie die entsprechende Graphik-Datei über Eine bestehende Bilddatei öffnen.

Klicken Sie nun auf `Weiter >`

Schritt 2 – Layout des Originalbildes

Abbildung 3.4: Schritt 2

Abbildung 3.4 Über dieses Fenster können Sie Omnipage mitteilen, auf welches Layout es sich konzentrieren soll. Besteht Ihr Dokument überwiegend aus Tabellen, so sollten Sie den dritten Punkt (Kalkulationstabelle) aktivieren, damit Omnipage möglichst viel von der vorliegenden Tabellenstruktur zu retten versucht.

Für den Fall, dass Sie sich nicht sicher sind, wählen Sie den letzten Punkt (Gemischt) und lassen Omnipage selbst entscheiden.

Schritt 3 – Festlegen von Sprache und Bearbeitungsaufwand

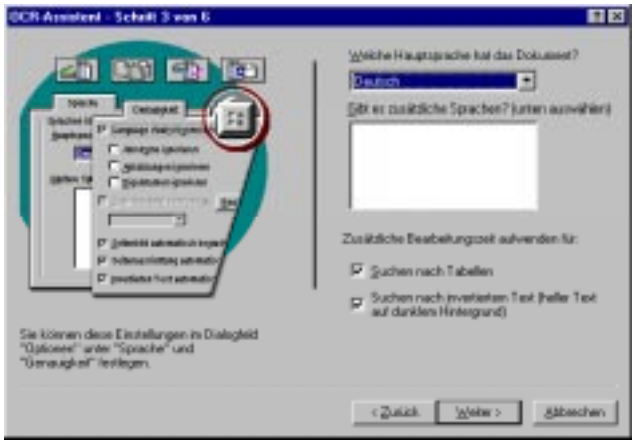
OCR-Assistent - Schritt 3 von 6
Welche Hauptsprache hat das Dokument?
Deutsch
Gibt es zusätzliche Sprachen? (unten auswählen)
Zusätzliche Bearbeitungszeit aufwenden für:
Suchen nach Tabellen
Suchen nach invertiertem Text (heller Text auf dunklem Hintergrund)
Sie können diese Einstellungen im Dialogfeld "Optionen" unter "Sprache" und "Genauigkeit" festlegen.
< Zurück
Weiter >
Abbrechen

Abbildung 3.5: Schritt 3

! → Im Standardumfang von Omnipage ist leider nur ein deutsches Wörterbuch enthalten. Da die Genauigkeit der Erkennung jedoch stark von dem zugeordneten Wörterbuch abhängt, ist es sinnvoll, bei der Interpretation fremdsprachiger Texte eventuell ein entsprechendes Wörterbuch hinzuzukaufen! (Der Softwarehändler Ihrer Wahl hilft Ihnen dabei gern weiter. . .)

Dies bedeutet nicht, dass ein Erkennen beispielsweise englischer Texte mit einem deutschen Wörterbuch unmöglich wäre, da das System eine korrekte Erkennung stets in Zusammenhang mit dem zugeordneten Wörterbuch durchführt, wird es die betreffende Stelle innerhalb Ihres Dokuments entsprechend markieren.

Durch die letzten beiden Optionen auf dieser Seite können Sie innerhalb eines Fließtextes noch verstärkt nach Tabellen suchen lassen – sofern Sie sich im letzten Schritt nicht direkt dazu durchgerungen haben, ein Tabellenlayout zu definieren – bzw. innerhalb des Dokuments vorhandene Graustufen auszugleichen.

Schritt 4 – Übernahme des Originallayouts

Die Übernahme des Originallayouts wird wohl am häufigsten gewünscht, bzw. von Einsteigerinnen und Einsteigern in aller Regel stillschweigend vorausgesetzt.

Trotzdem drängt sich die Frage auf, ob eine 1:1 Umsetzung eigentlich sinnvoll ist, da nach meinen persönlichen Erfahrungswerten eingescannte Texte häufig in bereits bestehende oder in der Erstellung befindliche Dokumente integriert werden sollen, was wiederum eine generelle Neu-Formatierung erzwingt.

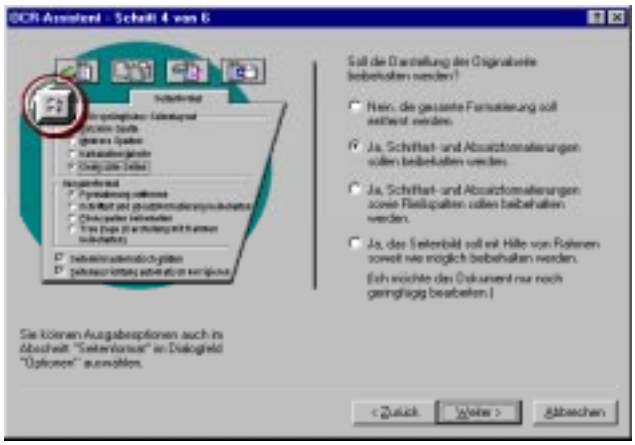

Abbildung 3.6: Schritt 4

Ein guter Mittelweg ist es, die Schrift- und Absatzformatierungen (bei mehrspaltigem Text auch das Spaltenlayout) zu übernehmen (Optionen 2 und 3 in Abbildung 3.6) – von einer Übernahme aller Formate, einschließlich Rahmen etc., ist im Regelfall jedoch eher abzuraten.

Schritt 5 – OCR-Ergebnisse prüfen

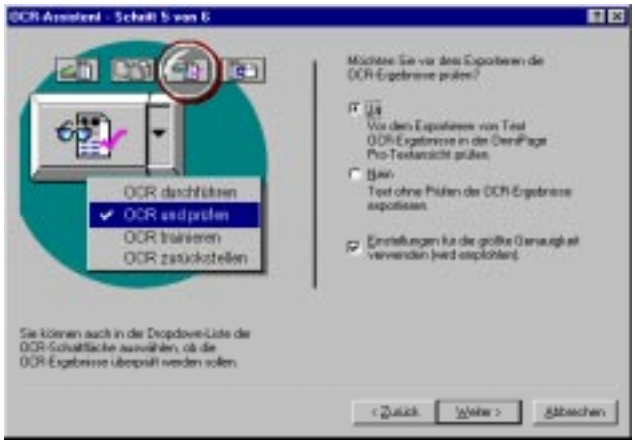

Abbildung 3.7: Schritt 5

Nach dem OCR-Lauf können Sie das Resultat direkt in Omnipage prüfen bzw. korrigieren. Hierzu startet das System eine Korrekturhilfe wie sie z.B. auch von der Rechtschreibkorrektur unter Word für Windows her bekannt ist.

Sofern Sie Omnipage vertrauen, oder Ihren Text im Anschluß direkt über Ihre Textverarbeitung kontrollieren, können Sie die Korrekturhilfe abschalten.

Schritt 6 – Dokument speichern

Abschließend muss Omnipage noch wissen, was letztendlich mit Ihrem Dokument geschehen soll. Mein Tipp: Speichern Sie es in einem Format, welches Ihre Textverarbeitung interpretieren kann, direkt auf die Festplatte.

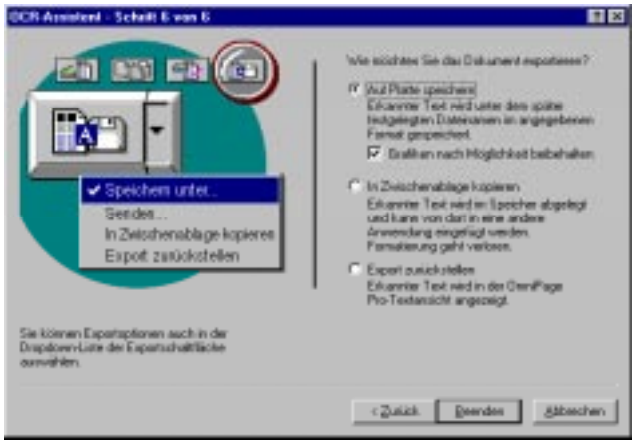

Abbildung 3.8: Schritt 6

Tipp! Sofern Sie die Option Grafiken nach Möglichkeit beibehalten aktivieren, werden innerhalb des Textes erkannte Bilder in eigenen Dateien zur Weiterverwendung gesichert.

Abbildung 3.9 Omnipage wird, sofern Sie die Option Speichern ausgewählt haben, nach der Verarbeitung das folgende Fenster auf dem Bildschirm aufbauen.

Abbildung 3.9: Schritt 6: Omnipage legt Ihr Dokument auf der Festplatte ab.

Tragen Sie in dem Fenster den von Ihnen gewünschten Dateinamen ein, und wählen Sie aus der Liste darunter das Format Ihrer Textverarbeitung.

3.3 Weitere Tipps und Kniffe

Über dem oben aufgezeigten Weg sind die Fähigkeiten von *Omnipage Pro* natürlich bei weitem nicht erschöpft; es soll Ihnen an dieser Stelle lediglich ein möglichst einfacher und schneller Weg von der Vorlage bis zum übernahmefertigen Dokument aufgezeigt werden.

Für die meisten Vorlagen wird die gezeigte Technik sicherlich ausreichen – für den Fall, dass Sie jedoch ein komplexeres Dokument „erkennen“ müssen, bzw. häufiger Dokumente umwandeln, und aus diesem Grund Ihre OCR-Läufe „planen“ sollten, sei an dieser Stelle ein Verweis auf das *Omnipage Pro*-Benutzungshandbuch gestattet, welches Sie ebenfalls über die Menüreihenfolge `Start - Programme - Omnipage 9.0 Prof.`[4] erreichen.

Auf 130 Seiten verteilt finden Sie im Benutzungshandbuch alles was Ihnen die Arbeit mit *Omnipage Pro* zusätzlich vereinfacht.

3.4 Benutzung des Scanners HP 9100C Digital Sender

Mit der oben aufgeführten Konfiguration des Rechners `Karin` plus anhängendem Scanner besitzen Sie eigentlich alle Möglichkeiten, um Ihre Vorlage in ein maschinenlesbares Dokument zu umzuwandeln.

Als Schwachpunkt dieser Konfiguration ist jedoch sicherlich der Scanner und dessen arg eingeschränkte Verarbeitungsgeschwindigkeit zu betrachten. Da der HP ScanJet 4c primär dazu beschafft wurde, hochauflösende farbige Graphiken zu erfassen, sind die Grenzen seiner Einzugsfähigkeit mit ca. *vier Seiten pro Minute* schnell erreicht.

Nehmen wir als Beispiel einmal an, dass Sie eine Kurseinheit von ungefähr 200 Seiten einscannen möchten, so sind Sie bereits eine knappe halbe Stunde mit dem Einlesen der Vorlage beschäftigt, ohne dass auch nur eine einzige Seite zu diesem Zeitpunkt konvertiert worden wäre.

Zur Steigerung dieses mageren Durchsatzes steht Ihnen im Raum der *URZ-Beratung* (Raum A003 im URZ) ein speziell zum Einlesen von Textdokumenten beschaffter Hochgeschwindigkeitsscanner zur Verfügung, der es auf eine Einlesegeschwindigkeit von ca. *15 Seiten pro Minute* bringt, was für unser obiges Beispiel bedeutet, dass das Erfassen des Dokuments bereits nach knapp sieben Minuten beendet ist.

Schritt 1: Starten des HP 9100C

! → *Vorabkontrolle*: Beim Hochfahren des PCs `Karin` sollte automatisch das Transferprogramm *HP Digital Sender Link* gestartet werden, was durch ein entsprechendes Symbol (in unserem Beispiel das erste von rechts) in der Windows *Symbolleiste* dargestellt wird.

[4] Oder, sofern Sie *Omnipage Pro* auf Ihrem eigenen Rechner installiert haben, unter `Start - Programme - Caere`

Sollte dieses Symbol nicht vorhanden sein, so starten Sie den Digital Sender Link bitte manuell über das Menü `Start - Programme - HP9100C - HP Digital Sender Link`.

Der nächste Schritt innerhalb unseres Verarbeitungsablaufs ist hier mehr organisatorisch zu betrachten – gehen Sie in den Raum der URZ-Beratung (Raum A003 im Gebäude AVZ I) und scannen Sie Ihre Vorlage ein.

Über das *Frontpanel* des Scanners legen Sie dabei fest, dass die eingescannten Seiten im Anschluss an den PC `Karin` gesendet werden.

! → Eine detaillierte Beschreibung der Bedienung dieses Scanners würde den Rahmen dieser Broschüre sprengen – sofern Sie sich daher unsicher sind, hilft Ihnen die anwesende Kollegin/der Kollege in der Beratung gern weiter. Die URZ-Beratung ist wochentags zwischen 9.00 Uhr morgens und abends 22.00 Uhr besetzt!

Schritt 2: Übertragung der eingescannten Seiten auf den PC `Karin`

Sobald Sie Ihre Seiten auf dem Scanner eingelesen haben, werden Sie zum PC `Karin` zurückkehren und wahrscheinlich feststellen, dass sich ich der Zwischenzeit bereits einiges getan hat.

Durch den aktiven Digital Sender Link (s.o.) wurde nach Übertragung der eingescannten Seiten das Programm *Adobe Circulate* gestartet, das Ihnen Ihr Dokument ikonisiert darstellt.

Abbildung 3.10

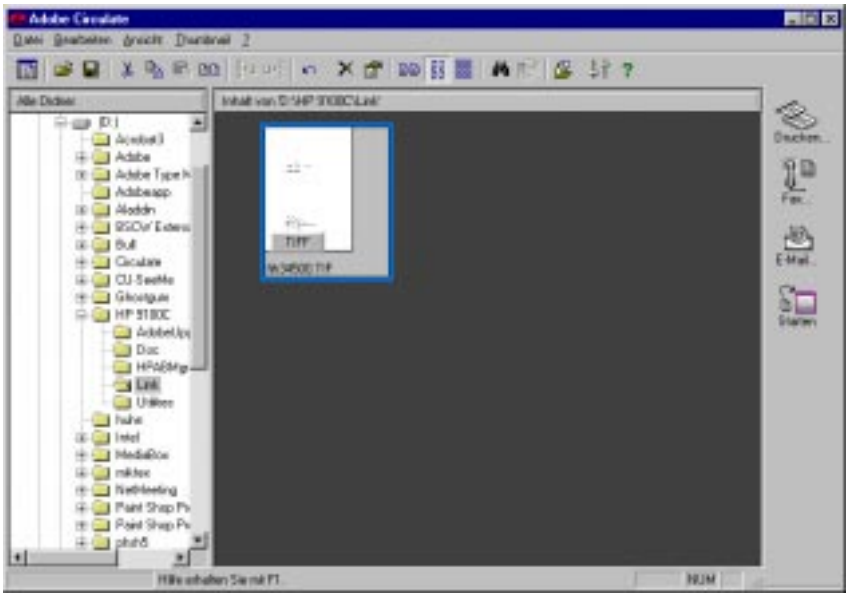

Abbildung 3.10: Überraschung! Der Scanner hat Ihr Dokument an *Circulate* durchgereicht.

So gehts weiter: Klicken Sie Ihr Dokument mit der *linken* Maustaste an, „ziehen“ es auf das Symbol `Starten` und danach, in dem erscheinenden Pop-Up-Fenster, auf den Eintrag `omnipage` – *Omnipage Pro* wird durch diesen Vorgang automatisch gestartet.

Abbildung 3.11 auf der nächsten Seite

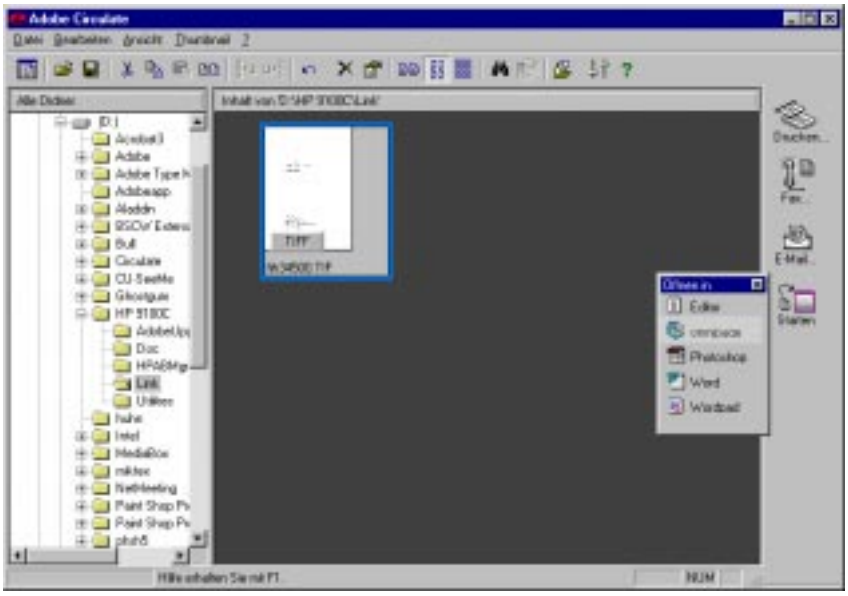

Abbildung 3.11: Weiter gehts in Richtung *Omnipage Pro.*

Alles weitere funktioniert wie im letzten Kapitel gezeigt, d.h. Sie aktivieren den *Omnipage Assistenten*, s.a. Kapitel 3.2, der Ihnen den Rest der Arbeit weitestgehend abnimmt.

! → Beim Start des Assistenten bemerkt dieser, dass bereits ein Dokument, nämlich das soeben übertragene, geladen ist und präsentiert Ihnen, zu Ihrer Sicherheit, das folgende Dialogfenster.

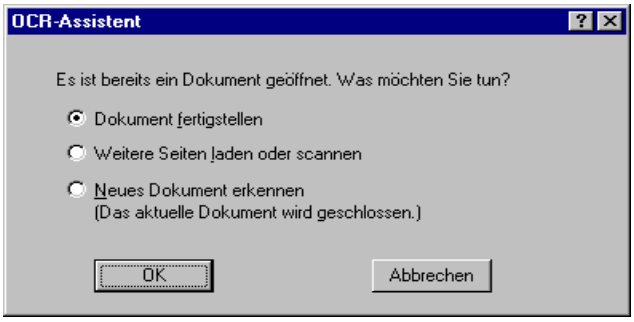

Abbildung 3.12: Der Assistent geht auf Nummer Sicher.

Bestätigen Sie hier unbedingt, dass Sie das geöffnete Dokument fertigstellen wollen.